

(12) UK Patent Application (19) GB (11) 2 330 930 (13) A

(43) Date of A Publication 05.05.1999

(21) Application No 9820556.0

(22) Date of Filing 21.09.1998

(30) Priority Data

(31) 08936336

(32) 24.09.1997

(33) US

(71) Applicant(s)

Ricoh Company Limited
(Incorporated in Japan)

3-6 Nakamagome 1-chome, Ohta-ku, Tokyo 143-8555,
Japan

(72) Inventor(s)

John Cullen

Jonathan J Hull

(74) Agent and/or Address for Service

J A Kemp & Co.

14 South Square, Gray's Inn, LONDON, WC1R 5LX,
United Kingdom

(51) INT CL⁶

G06F 17/30

(52) UK CL (Edition Q)

G4A AUSB

(56) Documents Cited

EP 0722145 A1

EP 0631245 A2

EP 0601759 A1

WO 95/12173 A1

(58) Field of Search

UK CL (Edition Q) G4A AUSB

INT CL⁶ G06F 17/30

(54) Abstract Title

Navigation system for document database

(57) An interactive database organization and searching system employs text search and image feature extraction to automatically group documents together by appearance. The system automatically determines visual characteristics of document images and collects documents together according to the relative similarity of their document images.

Fig.2A.

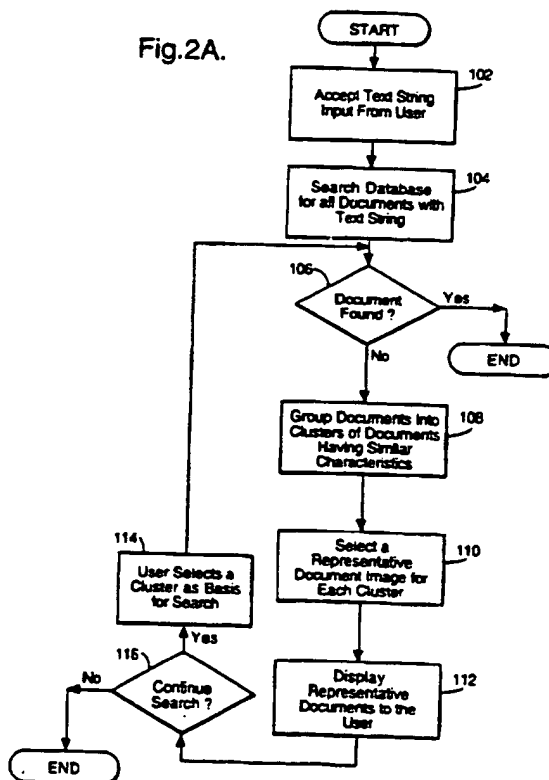


Fig. 1.

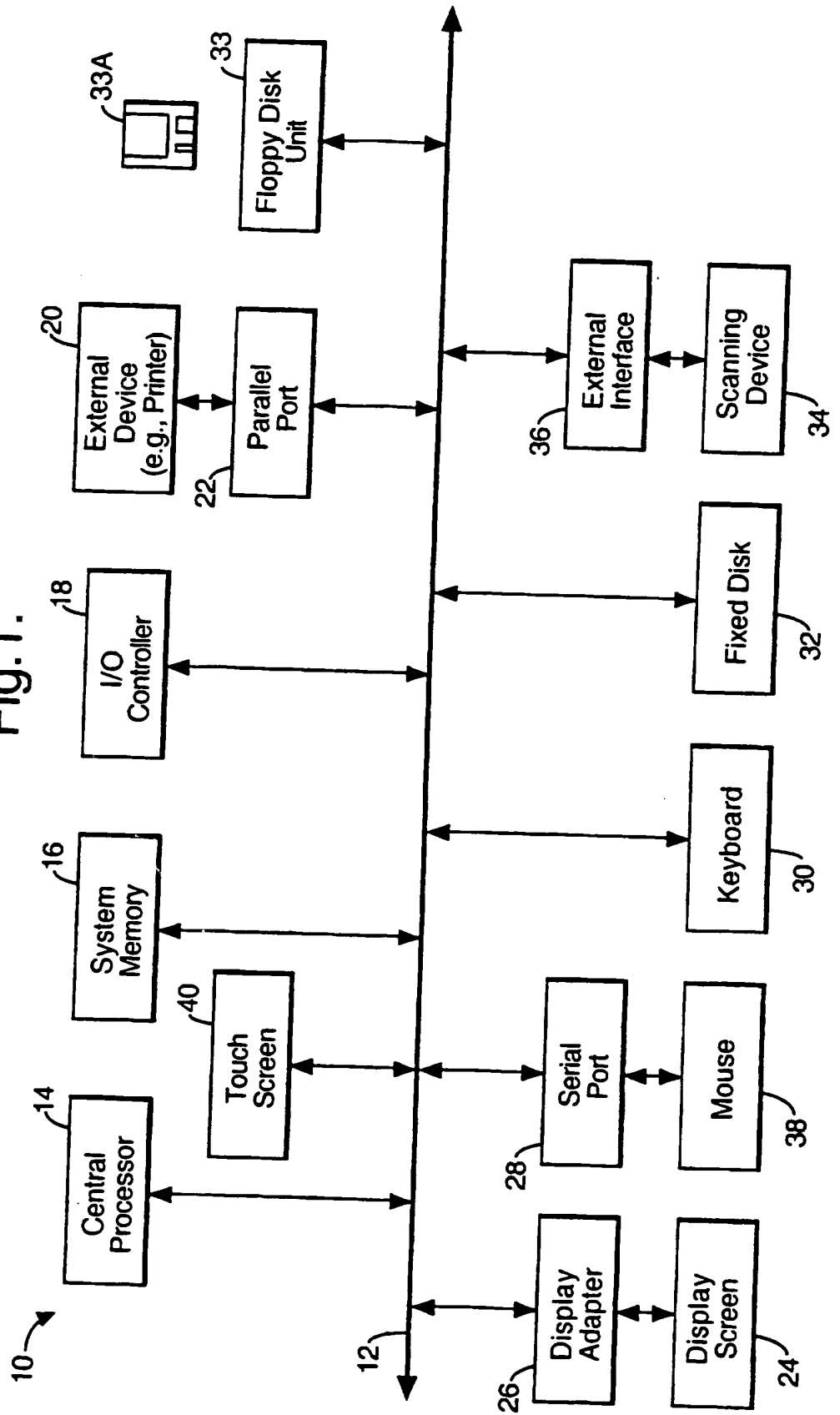


Fig.2A.

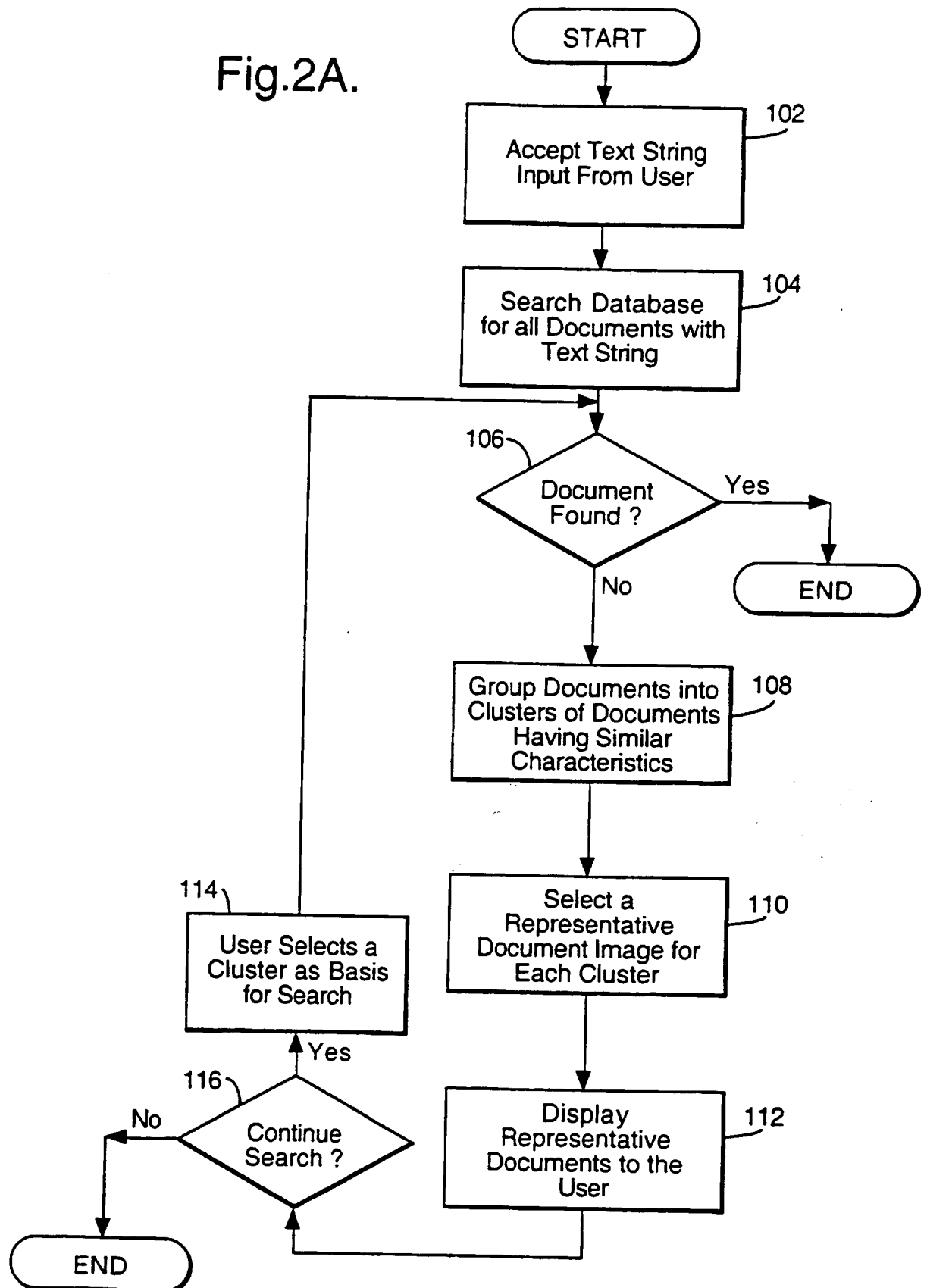


Fig.2B.

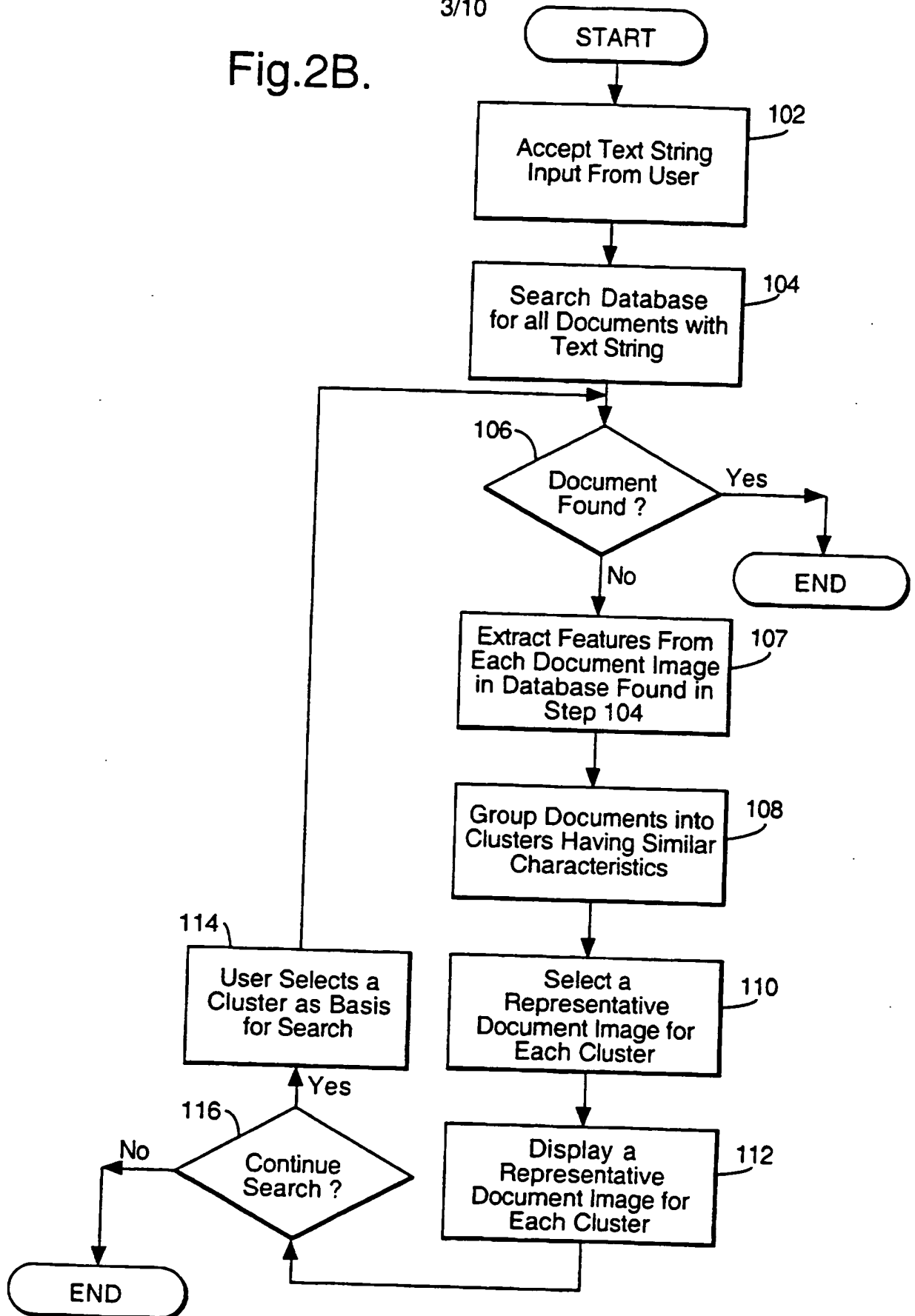


Fig.3A.

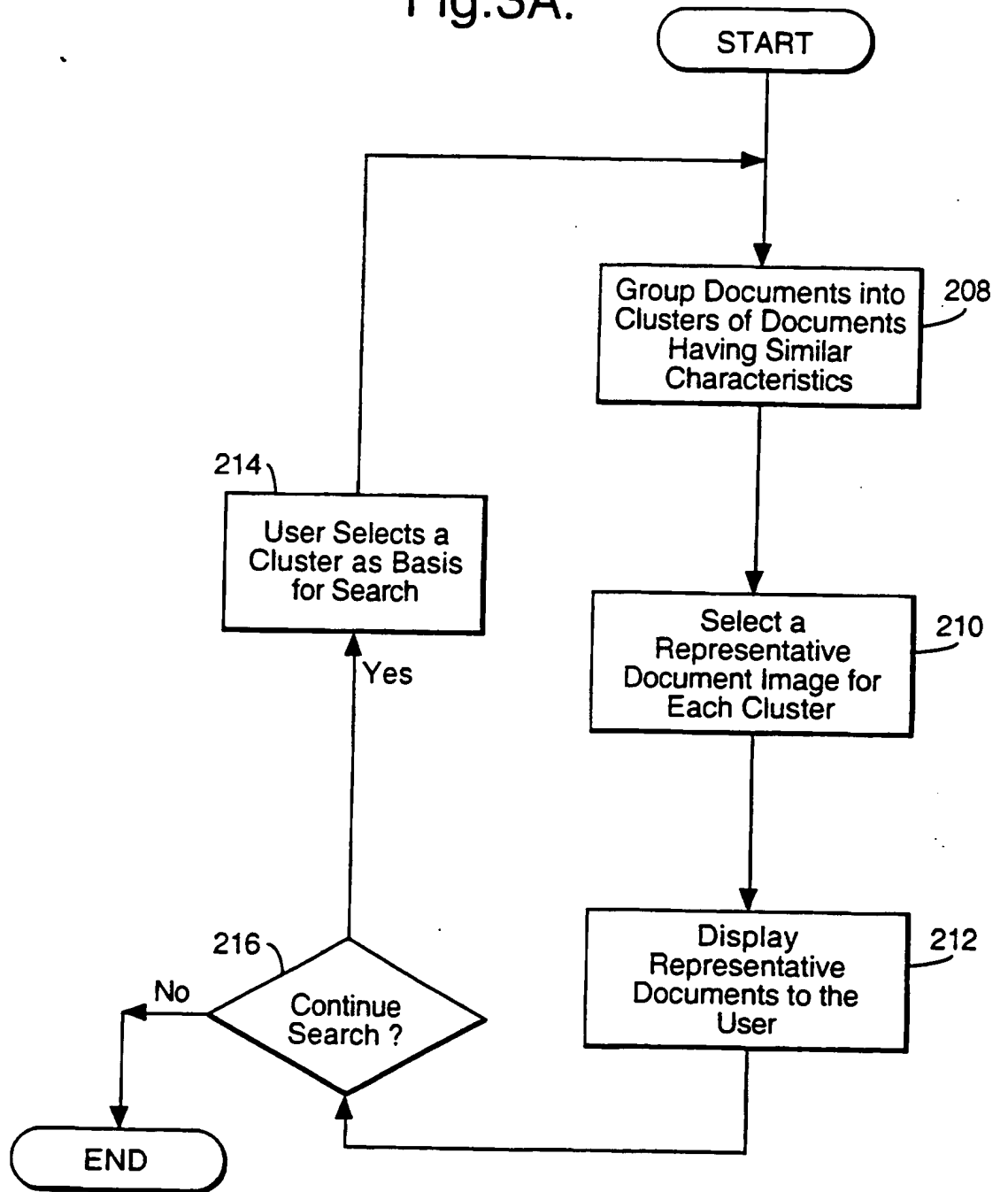


Fig.3B.

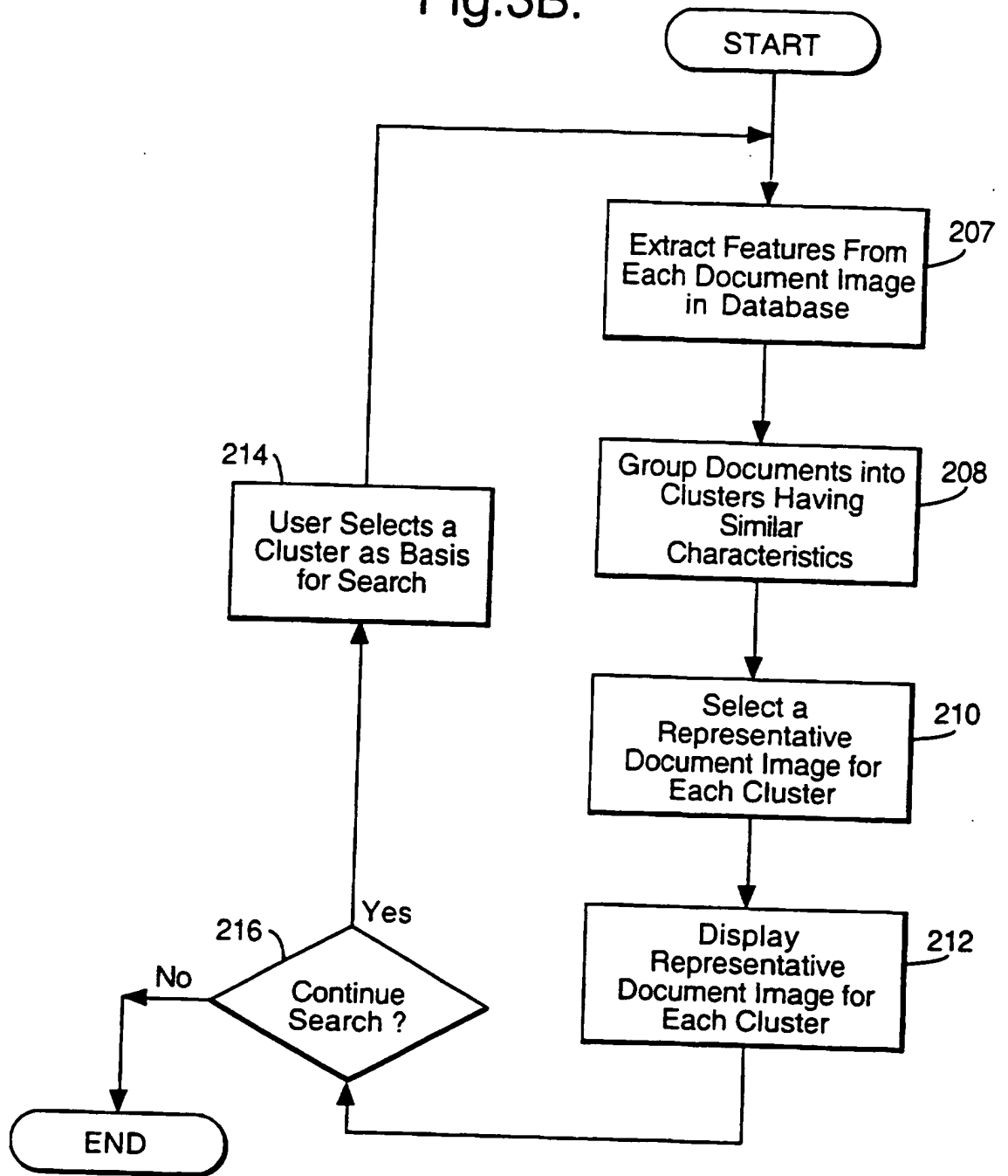


Fig.5.

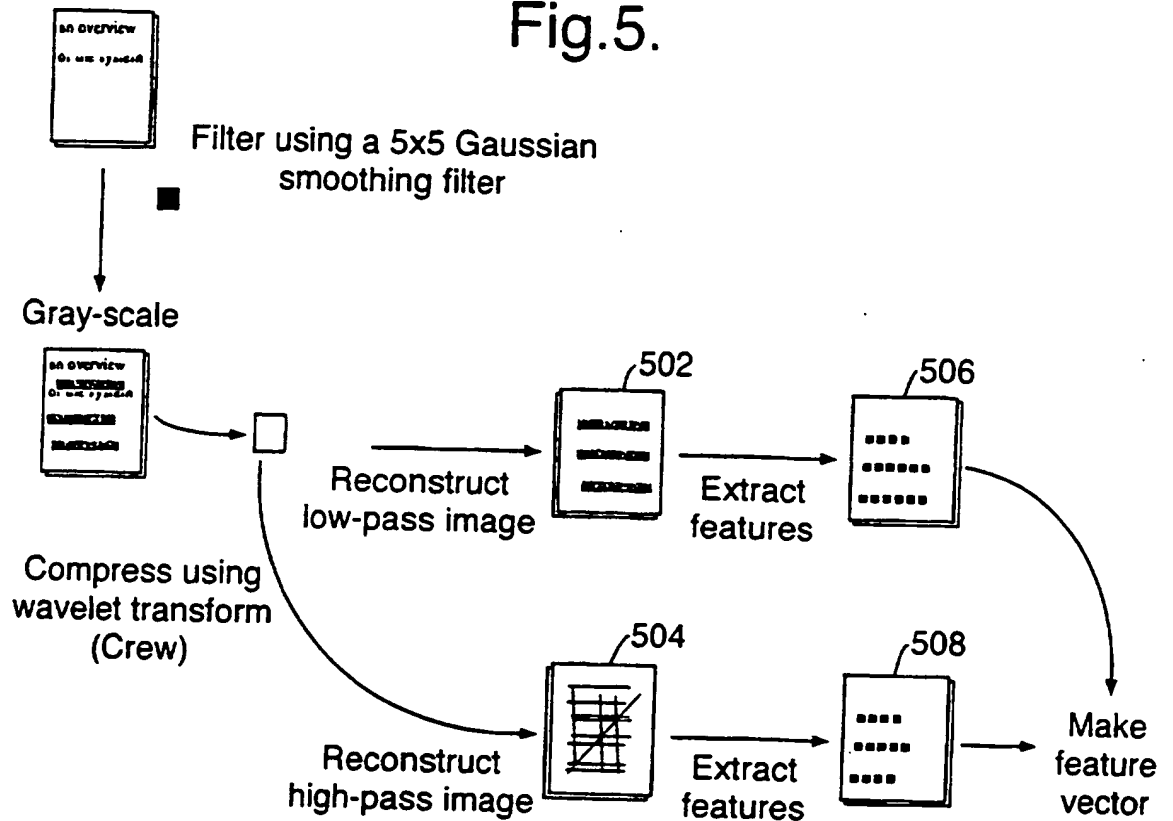


Fig.6.

Statistical moments ...

Consider a pixel values: $X_1 \dots\dots\dots X_N$

$$X_{mean} = \frac{1}{N} \sum_{j=1}^N X_j$$

$$X_{variance} = \frac{1}{(N-1)} \sum_{j=1}^N (X_j - X_{mean})^2$$

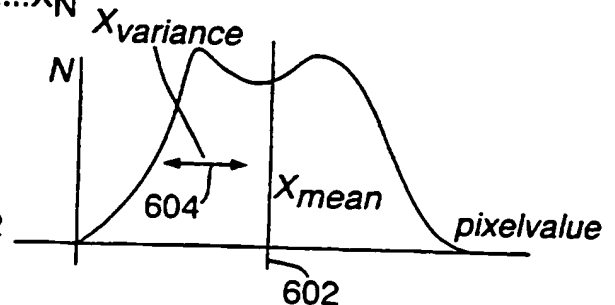
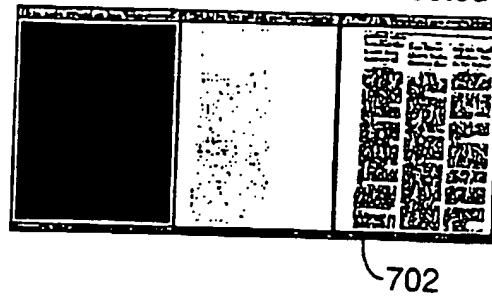


Fig.7.

Connected components



Number of words
and pictures

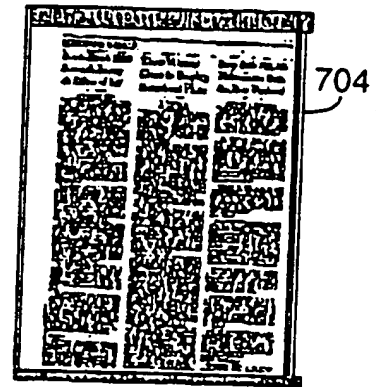
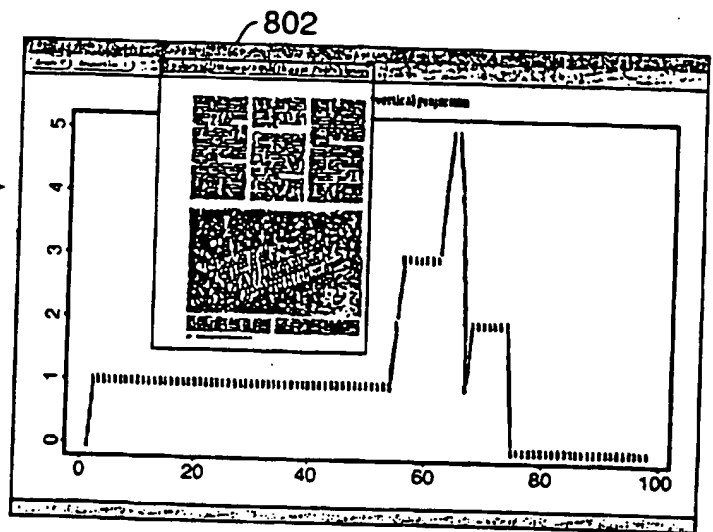
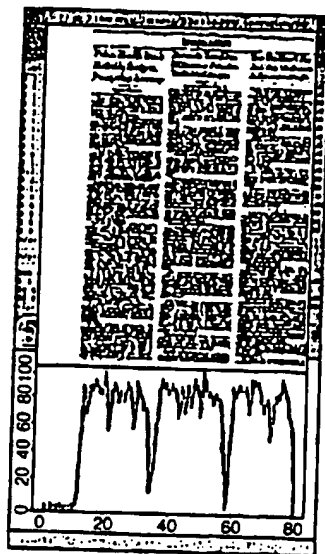


Fig.8.

Numbers of columns



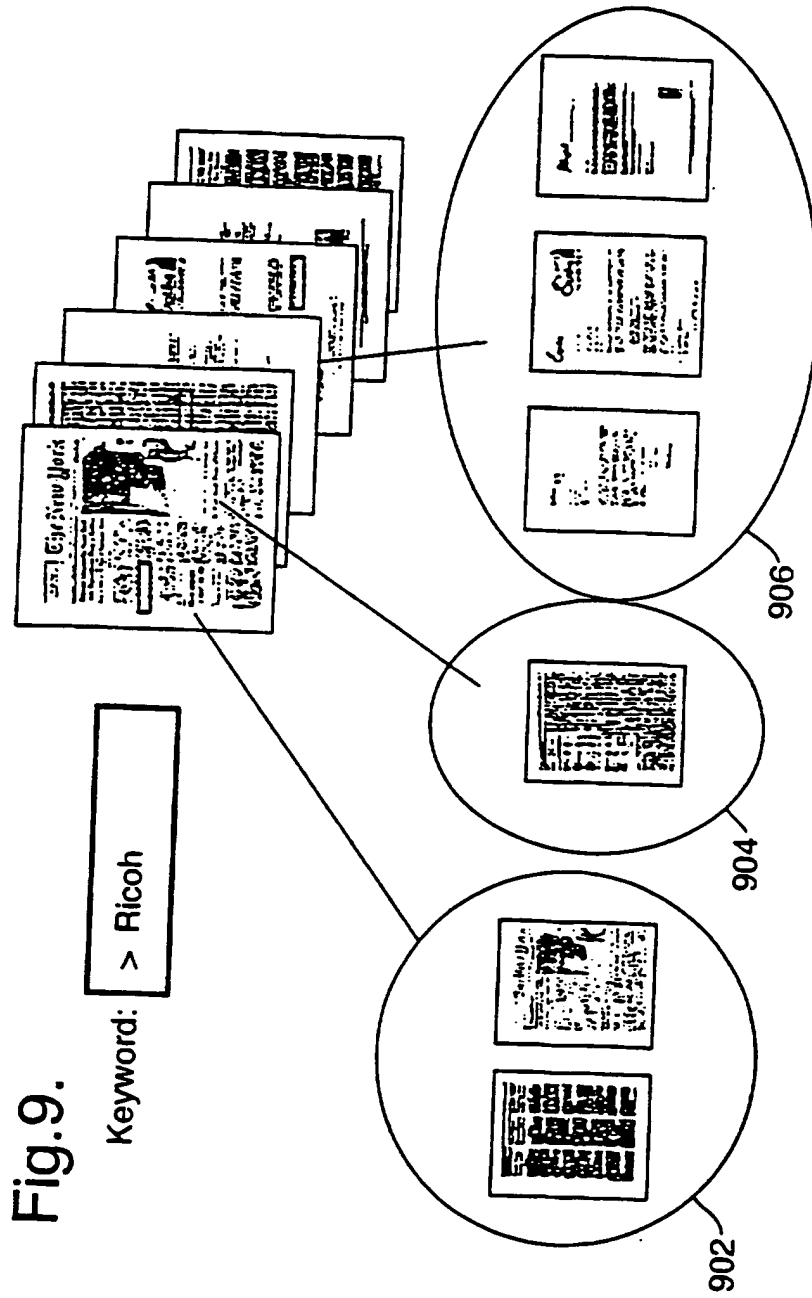
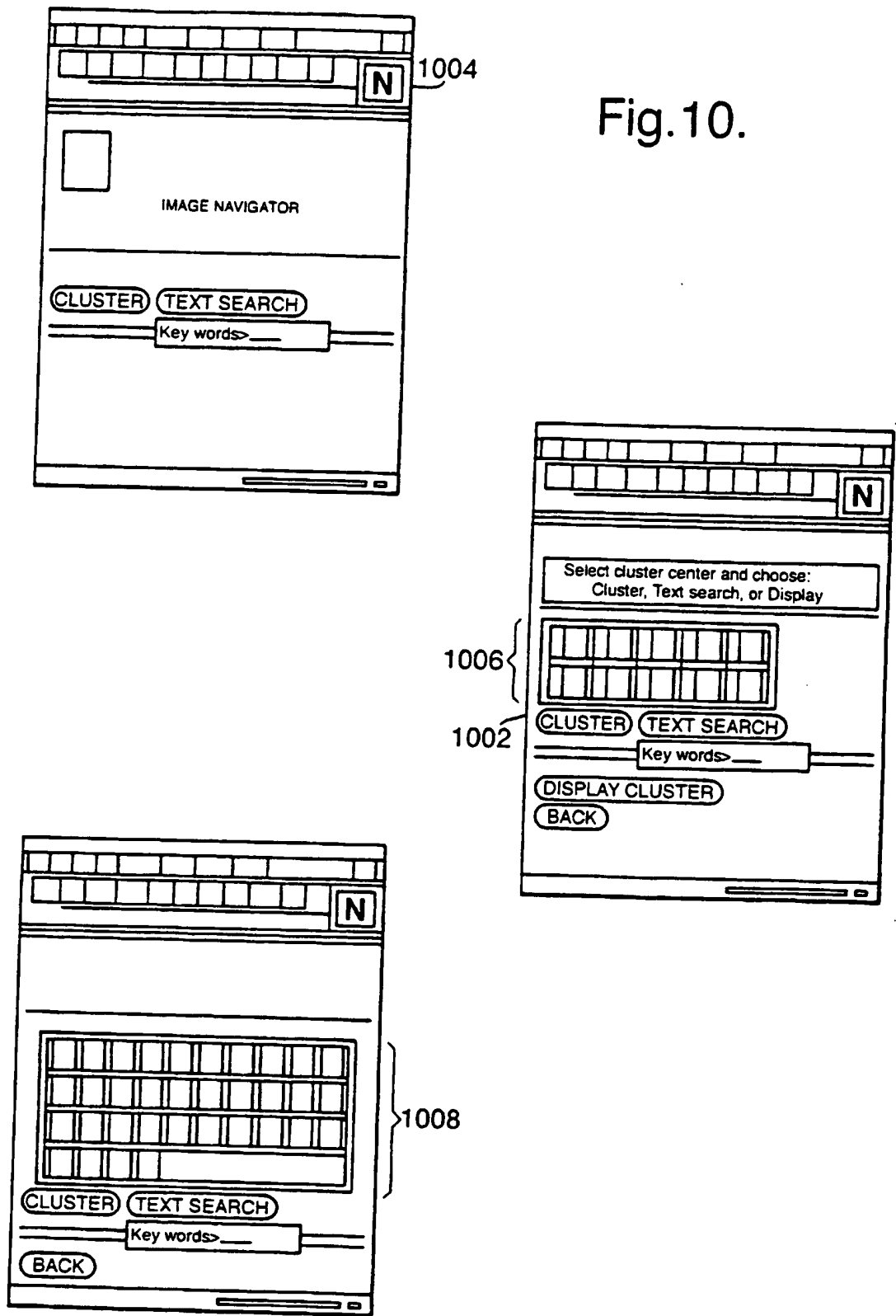


Fig.9.

Fig.10.



2330930

1 NAVIGATION SYSTEM FOR DOCUMENT IMAGE DATABASE

 This invention relates to document
management systems and more particularly to providing
a method of navigating through a database of document
5 images.

 The proliferation of low-cost, high-capacity
electronic storage of document images has enabled
users to keep ever increasing amounts and varieties of
documents, previously stored in hard copy format, as
10 electronic information online. While this revolution
in storage technology has reduced the cost of document
storage, it brings with it the need for more efficient
methods of searching through a myriad of online
documents to find a particular document or set of
15 documents of interest to the user.

 Methods for locating a document of interest
have been rudimentary at best. Typically, in these
methods, documents are scanned into the computer and
an Optical Character Recognition ("OCR") program
20 converts the image into a textual file. Next, a form
of keyword matching search is performed, with the
system either scanning the entire text of all
documents, or a set of carefully chosen keywords
thought to be representative of the document by a
25 person who initially classified the document.

1 The problem with the first approach is the high
search cost involved with traversing a large number of
documents in their entirety. The difficulty with the
second approach is that different persons will employ
5 different strategies to filing and retrieval. As the
heterogeneity of documents contained in databases
increases, the reliability of these traditional search
methods diminishes.

Recognizing the opportunity to exploit the
10 information content of the image portion of documents,
several attempts have been made to search for
documents based upon matching of small images
contained in the documents. For example, M.Y.
Jaisimha, A. Bruce and T. Nguyen in their work,
15 "DocBrowse: A system for Textual and Graphical
Querying on Degraded Document Image Data" describe a
system which searches for documents based upon company
logos in letterheads. D. Doermann, et. al. in
"Development of a General Framework for Intelligent
20 Document Retrieval," outline a system for matching
documents based upon generation and matching of an
image descriptor which describes low-level features
and high-level structure of a document.
Unfortunately, this method requires intensive
25 processing of the image information, which greatly

1 curtails its use in most commercial applications.

 While such methods provide document search
 capability via elemental matching of image
 characterization vectors, they do not provide a useful
5 method to organize a large database of document
 images. These and other shortcomings indicate that
 what is needed is a method and system for efficiently
 searching a database of document images. This method
 would expedite search by organizing the database
10 according to the textual as well as the visual
 characteristics of document images.

 The present invention provides an
15 interactive database organization and searching system
 which employs text search and image feature extraction
 to automatically group documents together by
 appearance. The system automatically determines
 visual characteristics of document images and collect
20 documents together according to the relative
 similarity of their document images.

 One representative embodiment is a method
 that includes the steps of accepting from the user a
 textual string keyword to serve as the basis for an
25 initial search, searching the textual component of

1 document images in the database for the keyword,
grouping document images having textual components
which contain the keyword into clusters of document
images based upon processing of the features extracted
5 from compressed representations of the document
images, displaying a representative document image for
each cluster of document images, and accepting input
from the user indicating a particular cluster of
document images upon which the system may perform
10 further search.

Another representative embodiment is a
method that includes the steps of grouping document
images together based upon feature information to form
clusters of document images having similar feature
15 characteristics, selecting a representative document
image from each cluster of document images, displaying
the representative document image to the user, and
accepting input from the user to select a cluster of
images upon which the system may perform further
20 search.

A related embodiment extracts feature
information from the compressed image prior to
performing the grouping steps described above.
Another related embodiment also permits the user to
25 specify a desired number of clusters.

1 The invention will be better understood by
reference to the following detailed description in
connection with the following drawings, in which:

5

Fig. 1 depicts a representative computer system suitable for implementing the invention.

Fig. 2A depicts a flowchart describing a representative querying operation of the database in a
10 preferable embodiment of the invention.

Fig. 2B depicts a flowchart describing a representative querying operation of the database in an alternative embodiment of the invention.

Fig. 3A depicts a flow chart of the steps
15 performed in organizing the database in an alternate preferable embodiment of this invention.

Fig. 3B depicts a flow chart of the steps performed in organizing the database in an alternate embodiment of this invention.

20 Fig. 4 depicts the use of compression on a document image to facilitate low cost storage and expedient manipulation of image components by the system.

Fig. 5 depicts extraction of image feature
25 information from compressed images to serves as the

1 basis for image grouping.

Fig. 6 depicts extraction of statistical moments from low frequency image information.

5 Fig. 7 depicts extraction of connected components of number of words and number of pictures from high frequency image information.

Fig. 8 depicts extraction of connected components of number of columns from high frequency image information.

10 Fig. 9 depicts the clustering of document images aspect of the invention.

Fig. 10 depicts the use of a web browser to implement the display and user interface portions of the invention.

15

Representative System Suited to Practice the Invention

20 In a typical installation, the invention will be practiced on a computer system with the basic subsystems such as depicted in Fig. 1. In the representative system of Fig. 1, a computer system 10 includes bus 12 which interconnects major subsystems
25 such as central processor 14, system memory 16,

1 input/output (I/O) controller 18, an external device
 such as a printer 20 via parallel port 22, display
 screen 24 via display adapter 26, serial port 28,
 keyboard 30, fixed disk drive 32 and floppy disk drive
5 33 operative to receive a floppy disk 33A. Many other
 devices can be connected such as scanning device 34
 connected via external interface 36, mouse 38
 connected via serial port 28 and touch screen 40
 connected directly. Many other devices or subsystems
10 (not show) may be connected in a similar manner.
 Also, it is not necessary for all of the devices shown
 in Fig. 1 to be present to practice the present
 invention, as discussed below. The devices and
 subsystems may be interconnected in different ways
15 from that shown in Fig. 1 without impairing the
 operation of the system. The operation of a computer
 system such as that shown in Fig. 1 is readily known
 in the art and is not discussed in detail in the
 present application. Source code to implement the
20 present invention may be operably disposed in system
 memory 16 or stored on storage media such as fixed
 disk 32 or floppy disk 33A. An image database may
 also be stored on fixed disk 32.

 Display screen 24 is similar to that in use
25 on standard computers such as personal computers,

1 workstations or mainframe computers employing a CRT
screen or monitor. Various forms of user input
devices may be used with the present invention. For
example, a mouse input device 38 that allows a user to
5 move a pointer displayed on the display screen in
accordance with user hand movements in a standard user
input device. A mouse usually includes one or more
buttons on its surface so that the user may point to
an object on the screen by moving the mouse and may
10 select the object, or otherwise activate the object,
by depressing one or more buttons on the mouse.
Alternatively, a touch screen allows a user to point
to objects on the screen to select an object and to
move the selected object by pointing to a second
15 position on the screen. Various buttons and controls
may be displayed on the screen for activation by using
the mouse or touch screen. Fixed disk drive 32 may be
a hard disk drive or an optical drive or any medium
suitable for storing a database of document images.

20

Overview of Querying Operation Performed on the
Image Database

One unique and innovative feature of this
invention is the intuitive manner in which image based
25 search can be conducted on the documents in the

1 database without requiring the user to build a
representative document image as is required in the
art. See Japanese Laid-Open Patent Application No. 9-
237282, which corresponds to U.S. Patent Application
5 S.N. 08/431,059, entitled, "Image Database Browsing
and Query Using Texture Analysis".

In one particular embodiment of the
invention depicted in Fig. 2A, the textual portion of
documents is extracted from document images using OCR
10 scanning and is made available in the database. This
database can reside in any or multiple storage media
in the computer system, such as the fixed disk 32 or
system memory 16. The search procedure commences with
an initial query step 102, in which the user inputs
15 one or more keywords (i.e. text string or a
combination of text strings) into the system via an
input device such as terminal keyboard 30 and display
screen 24.

Searching proceeds for the text string in
20 the textual portions of the documents contained in the
database in text based search step 104. If the text
based search yields the document of interest, then the
user may discontinue any further processing in step
106. Otherwise, documents containing the text string
25 become the basis for image based search. In a

1 preferred embodiment, features extracted from
compressed representations of document images are
available in the database. Documents meeting the text
based search are grouped together in grouping step 108
5 to form clusters based upon similarity of the features
extracted from each document image. A representative
document image is selected in step 110 for each
cluster of document images formed in the grouping step
108.

10 The representative document images are
displayed to the user on the display 24 in display
step 112. In a preferable embodiment, as depicted in
Fig. 4, a compressed representation of each
representative document image is displayed as an icon
15 402 using a web browser 404 as a user interface. A
related embodiment displays uncompressed
representative images. In the preferable embodiment,
the user may select a particular cluster as the basis
for further search by indicating to the system, with a
20 mouse 38 or other input device, the representative
document icon for the cluster which is to be the basis
of further search as in step 114.

Search proceeds by applying the grouping
step to the selected cluster of documents, sub-
25 dividing this cluster into a new set of clusters 108,

1 each having a new representative document image.
Search continues until either the document is found
106, or the user chooses to discontinue the search
116.

5 In a related embodiment depicted in Fig. 2B,
a compressed representation of each document image is
made available in the database. Features are
extracted from these compressed representations in an
extraction step 107, interposed between document
10 complete step 106 and grouping step 108. Search
proceeds with grouping step 108 as in the prior
embodiment.

Another related embodiment also includes the
step of permitting the user to choose the number of
15 clusters desired before the first grouping step 108,
or at any time while navigating the database prior to
a new application of the grouping step.

An alternative embodiment of the invention
is depicted in Fig. 3A. As in the embodiment
20 discussed above features extracted from compressed
representations of document images are preferably
available in the database. Documents are grouped
together to form clusters in grouping step 208 based
upon similarity of the features extracted from each
25 document image. A representative document image is

1 selected in step 210 for each cluster of document
images formed in the grouping step 208. The
representative document images are displayed to the
user on the display 24 in display step 212. In a
5 preferable embodiment, the system accepts from the
user input parameters upon which further search may
continue in step 214. Search continues until the user
chooses to discontinue 216.

10 In a related embodiment depicted in Fig. 3B,
a compressed representation of each document image is
made available in the database. Features are
extracted from these compressed representations in an
extraction step 207 before grouping step 208. Search
proceeds with grouping step 208 as in the prior
15 embodiment.

Another related embodiment also includes the
step of permitting the user to choose the number of
clusters desired before the first grouping step 208,
or at any time while navigating the database prior to
20 a new application of the grouping step.

Figs. 4-8 demonstrate the use of image
processing techniques of compression, Fig. 4; features
extraction, Fig. 5, 6, 7 and 8; and grouping, Fig. 9
which form the basis of the search technique.

25 Fig. 10 depicts the use of a web browser to

1 implement the display and user interface portions of
the invention.

Compression Techniques

5 Compression reduces the cost of storing
large quantities of document images in a database.
Each document image is compressed via a compression
technique such as wavelet compression (see e.g. IEEL
Data Compression Conference, CREW: Compression with
10 Reversible Embedded Wavelets, March 1995, which is
incorporated herein by reference in its entirety for
all purposes), or other techniques for compression
known in the art.

 Fig. 4 is illustrative of wavelet
15 compression, which operates by recursively applying a
pyramidal transform to the image data 402, dividing
the image into high frequency information 404 and low
frequency information 406. CREW has several
advantages in this application. It decomposes an
20 image into high and low pass components relatively
quickly. It gives a lossy compression of 20:1, with
minimal noticeable image degradation. Finally, it
produces a low-pass sub-band image 412 in the upper
left hand corner of the low-frequency quadrant. The
25 sub-band image provides a recognizable iconic

1 representation of the document. This visually
recognizable version of the document can be
efficiently accessed and is useful as an index to the
document information.

5 Fig. 5 is illustrative of the splitting off
of low frequency image information 502 and high
frequency image information 504 in one embodiment of
the invention to extract different image features 506,
508. A document image in binary format 500 is acted
10 upon by a 5 x 5 Gaussian smoothing filter in
processing step 501 to yield a grayscale
representation of the document image 502. Wavelength
compression algorithm 503 is used to transform
grayscale representation 502 into a compressed
15 representation 504. Low-pass filter step 505
separates low frequency image information 506 from the
compressed image representation 504. Analogously,
high-pass filter step 509 separates high frequency
image information 510 from the compressed image
20 representation 504. Features extraction step 507
performed on the low frequency image information 506
yields the mean pixel value and the variance of the
pixel values features 508. Analogously features
extraction step 511 performed on the high frequency
25 image information 510 yields the number of words,

1 number of pictures, and number of columns features
512. The low frequency and high frequency feature
information is amalgamated together in step 513 to
produce a feature vector 514.

5 In one embodiment of the invention, the low
pass component of an image 410 is used as an icon for
indexing into document image databases.

Feature Extraction

10 Feature extraction on the low frequency
image information resulting from the compression
yields the statistical moments of the image pixel
values as depicted in Fig. 6. In one particular
embodiment of the invention, the mean 602 and variance
15 604 of the pixel values are the statistical moments
which are calculated from low frequency information
according to the following formulas:

$$X_{mean} = \frac{1}{N} \sum X_j$$

20 and,

$$X_{variance} = \frac{1}{(N-1)} \sum (X_j - X_{mean})^2$$

wherein X_j is the value of each pixel in the
low frequency document image.

25 Figs. 7 and 8 are illustrative of features

1 extraction applied to the high frequency image
information resulting from the compression step.
Connected components are extracted. In one particular
embodiment of the invention, the features of total
5 number of text words 702, pictures 704 and text
columns 802, are extracted from the high frequency
image information. Features extraction on the high
frequency image information is done by looking for
connected components. The first step in processing is
10 to perform a histogram equalization, in which the
minimum and maximum gray values are calculated, then
their range is adjusted to have values between 0 and
255. Histogram equalization is a standard image
procedure technique well known to persons of ordinary
15 skill in the art. Finally, a connected component
algorithm, which in the preferred embodiment is a
four-connected component algorithm, is applied to the
image information.

According to a four-connected component
20 algorithm, processing of image data is done by looking
at the four sides of a particular pixel for other
pixels of similar gray level in searching for
connected components. Pixels adjacent to the pixel
under study at any of the four sites are aggregated
25 together to form a connected component. By contrast,

1 an eight-connected component algorithm would look not
only to the four sides of a pixel, but also to pixels
adjacent at any of the four vertices of a pixel in
locating connected components.

5 Once connected components have been
identified, features such as the total number of text
words 702, pictures 704 and text columns 802, may be
extracted from the connected component information.
The feature of total number of text words is
10 determined by examining the total number of connected
components below a certain threshold size. The
threshold value is set to discern connected components
which belong to text as from connected component
regions which are associated with pictures. The count
15 of connected components below the threshold is the
number of words. The count of the connected
components exceeding the threshold is the number of
pictures.

Fig. 8 depicts the processing underlying the
20 determination of the number of columns of text. A
plot of connected components (y-axis) vs. locations
(x-axis) is depicted in Fig. 8 graph 804. Plot 802 in
Fig. 8 depicts the number of transitions in graph 804
(y-axis) vs. the number of connected components (x-
25 axis). In the plot designated by 802, the number of

1 transitions is indicative of the number of columns in
the image.

From observing the dominant plateaus in
graph 802, it is possible to determine the predominant
5 number of columns in a document. In the case of 802,
the large picture causes this to be a predominantly
one column document. However, the second plateau
indicates that three columns are also present in the
document. The peak detected at some high number
10 indicates the word boundary noise, beyond this point
no further column information can be extracted.

Grouping Document Images into Clusters

Document images are grouped together based
15 upon feature information to form clusters of document
images, each cluster containing document images having
similar feature characteristics. One method of
performing grouping is to employ a k-means algorithm.
In one embodiment of the invention depicted in Fig. 9,
20 grouping of images is performed into a selected number
of groups 902, 904, 905. The desired number of groups
may be chosen by the user, based upon the user's
preference for granularity in the search vs the number
of groups the user can track without promoting
25 confusion. Grouping can be performed recursively on

1 the documents within any group, forming a hierarchical
organization in the database.

Displaying a Representative Document Image

5 A representative document image for display
to the user is automatically selected by the system
from each cluster of documents images. In one
embodiment of the invention, a center is calculated
and the nearest image to that center is labeled as the
10 characteristic image for each cluster generated by the
grouping step. In a preferred embodiment of the
invention, a web browser is used as the user interface
for displaying document image representations and
permitting the user to indicate which document image
15 cluster should serve as the basis of further
searching. As depicted in Fig. 10, a compressed
representation of each representative document image
is displayed as an icon 1002 using a web browser 1004
as a user interface. A related embodiment displays
20 non-compressed representative images. In a preferable
embodiment, the user may select a particular cluster
as the basis for further search by indicating to the
system, with a mouse 38 or other input device, the
representative document icon for the cluster which is
25 to be the basis of further search 1006. Search

1 continues by applying the grouping step to the
selected cluster of documents, sub-dividing this
cluster into a new set of clusters 1008, each having a
new representative document image. In an alternative
5 embodiment non-compressed document images are
displayed.

Searching the World Wide Web for Document Images

In a related embodiment, documents
10 retrieved from World Wide Web search engines, e.g.,
Altavista or Infoseek, may be browsed using methods
according to the invention. Users typically receive a
multiplicity of documents from a multiplicity of
different sources returned in response to a simple
15 text-based query. It is difficult to determine which
documents are actually of interest. However, users
desirous of a particular type of document
distinguishable by its visual appearance, e.g., a
scientific paper that contains mostly two-column text,
20 can quickly obtain only those documents using
techniques according to the invention.

The invention has now been explained with
reference to specific embodiments. Other embodiments
will be apparent to those of ordinary skill in the
25 art. It is therefore not intended that the invention

1 be limited, except as indicated by the appended
claims.

5

10

15

20

25

1 WHAT IS CLAIMED IS

5

1. A method for searching a database containing a plurality of document images, each document image having a textual component, a compressed representation, and a non-compressed representation, for a particular document image, said
10 method comprising the steps of:

accepting text from a user as a keyword to search;

searching the textual component of said
15 document images for said keyword;

grouping document images having textual components which contain said keyword into a plurality of clusters of document images based upon processing of the compressed representation, or the non-
20 compressed representation of the document images;

displaying, based on said processing, a representative document image for each cluster of the plurality of clusters of document images; and

accepting input from the user indicating a
25 particular cluster of document images.

2. A method according to claim 1 wherein said compressed representation is generated from said document image by applying a CREW algorithm to said document image.

3. A method according to claim 1 or 2 wherein said processing of the image component in the grouping step comprises the steps of:

10 extracting image feature information about said particular document image; and
applying a clustering algorithm to said image feature information to form clusters of document images.

15 4. A method according to claim 3 wherein said clustering algorithm comprises a "k-means" clustering algorithm to form a plurality of clusters of document images.

20 5. A method according to claim 4 wherein said plurality of clusters consists of between 5 and 10 document images.

25 6. A method according to claim 3, 4 or 5 wherein said extracting step comprises computing statistical information for said particular document image.

7. A method according to claim 6 wherein said extracting step additionally comprises computing component connections for said particular document image.

30 8. A method according to any one of the preceding claims wherein said processing in the displaying step comprises calculating a center for each cluster of document images and selecting the nearest document image
35 to said center as the representative document image.

9. A method according to any one of the preceding claims further comprising:

5 recursively applying the grouping, displaying and accepting input steps to form a hierarchical search pattern through the database.

10. A method according to any one of the preceding claims wherein said displaying step further comprises:

10 displaying said compressed representation of each representative document image using a web browser.

11. A method according to any one of the preceding claims wherein said displaying step further comprises:

15 displaying said representative document image using a web browser.

12. A method for organizing a plurality of document images in a database comprising the steps of:

20 compressing each particular document image in said plurality of document images;
 extracting image feature information about said particular document image;
 grouping said particular document images together to
25 form clusters of document images;
 selecting, based on processing, a representative document image for each particular cluster of document images; and
 displaying each particular representative document
30 image.

13. A method according to claim 12 wherein said compressing step comprises applying a CREW algorithm to said particular document image.

14. A method according to claim 12 or 13 wherein said extracting step comprises computing statistical information from said particular document image and
5 extracting text keywords from textual information contained in said particular document image.
15. A method according to claim 14 wherein said extracting step additionally comprises computing
10 component connections for said particular document image.
16. A method according to any one of claims 12 to 15 wherein said grouping step comprises a "k-means" clustering algorithm to form a plurality of
15 representative document groups.
17. A method according to claim 16 wherein said plurality of clusters consists of between 5 and 10 document images.
20
18. A method according to any one of claims 12 to 17 wherein said processing in the selecting step comprises calculating a center for each cluster of document images and selecting the nearest document image to said center
25 as the representative document image.
19. A method according to any one of claims 12 to 18 additionally comprising the step of accepting user input selecting a particular representative document image as a
30 starting point for recursive application of the grouping, selecting, displaying and accepting to form a hierarchical search method.
20. A method according to any one of claims 12 to 18
35 wherein said displaying step comprises displaying said

compressed representation of each representative document image using a web browser.

5 21. A method according to any one of claims 12 to 19 wherein said displaying step comprises displaying said representative document image using a web browser.

10 22. A computer program product comprising:
code that accepts text from a user as a keyword to search;
code that searches a database of document images, each document image having a textual component and a compressed representation, for document images with
15 textual components that contain said keyword;
code that groups document images together into clusters of document images based upon processing of the image component of said document images;
code that selects a representative document image
20 for each cluster for display;
code that displays said representative document images selected; and
a computer readable storage medium for storing the codes.

25 23. The computer program product of claim 22 further comprising:
code that generates said compressed representation from said document image by applying a CREW algorithm to
30 said document image.

24. The computer program product of claim 22 further comprising:
code that accepts user input selecting images from
35 the database from the representative document images displayed for recursive search forming a hierarchical

organization of said document images.

5 25. The computer program product of claim 22 wherein
said code that groups document images further comprises:
code that clusters document images into a plurality
of clusters employing a "k-means" algorithm.

10 26. The computer program product of claim 25 wherein
said plurality of clusters consists of between 5 and 10
document images.

15 27. The computer program product of claim 22 wherein
said processing of the image component of said code that
groups document images further comprises:
code that extracts image feature information from
said image component of said document images.

20 28. The computer program product of claim 27 wherein
said code that extracts image feature information further
comprises code that computes statistical information and
code that extracts connected component information.

25 29. The computer program product of claim 22 wherein
said code that selects a representative document image
comprises code that calculates a center for each cluster
of document images and selects the nearest document image
to said center as the representative document image.

30 30. The computer program product of claim 22 wherein
said code that displays said representative document
images selected comprises code that displays said
compressed representations of said document images using
a web browser interface.

35

31. The computer program product of claim 22 wherein
said code that displays said representative document
images selected comprises code that displays said
5 document images using a web browser interface.

32. A document image database organizing system
comprising:

an electronic storage unit that stores a document
10 image database;
a display that displays document images;
a processor unit coupled to said electronic storage
device and said display, said processor unit operative
to:
15 compress document images;
extract image feature information about document
images;
group document images together according to said
image feature information extracted;
20 select a representative document image for each
group formulated;
display said representative document image to a
user; and
accept from said user commands to manipulate
25 document images.

33. A method of searching a database substantially as
hereinbefore described with reference to the accompanying
drawings.

30

34. A computer system for storing and searching a
database constructed and arranged to operate
substantially as hereinbefore described with reference to
the accompanying drawings.

35



Application No: GB 9820556.0
Claims searched: 1,12,32

Examiner: Leslie Middleton
Date of search: 25 February 1999

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK CI (Ed.Q): G4A (AUIDB)

Int CI (Ed.6): G06F 17/30

Other:

Documents considered to be relevant:

Category	Identity of document and relevant passage	Relevant to claims
A	WO 95/12173 A3 (Teltech RNC)	
A	EP 0722145 A1 (IBM)	
A	EP 0631245 A2 (Xerox Corpn.)	
A	EP 0601759 A1 (Xerox Corpn.)	

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☒ **BLACK BORDERS**
- ☒ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☒ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☒ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

This Page Blank (uspto)